Case Retrieval for CBR based on Clustering to Support Decision

Abdelhak Mansoul
University of Skikda
Algeria
mansoul21@gmail.com

Baghdad Atmani
University of Oran 1 Ahmed Ben Bella
Algeria
atmani.baghdad@gmail.com

ABSTRACT: *Case-based reasoning has been widely adopted for decision support. A major operation in the CBR is the retrieval of similar cases. However, this operation has some weaknesses. One among them is the retrieval of several similar cases. On the other hand, sequentially processing all cases with a similarity has a complexity in presence of a great case base and several features. So, improving retrieval has been focused by a considerable amount studies using sequential calculation, non-sequential indexing, classification algorithms and Nearest Neighbor matching, etc., while others use hybridization of CBR with other reasoning methodologies. In this paper, we propose a novel approach based on CBR and clustering to improve the retrieval operation, and impacts positively the whole reasoning process. Our aim is to propose an available strategy for the retrieval task and also a valid decision support model. Finally, we present preliminary results and suggestions to extend our approach.*

Keywords: Decision Support, Case-Based Reasoning, Clustering, Data mining, CBR

Received: 23 September 2019, Revised 2 December 2019, Accepted 12 December 2019

DOI: 10.6025/jnt/2020/11/1/9-24

1. Introduction

Case-Based Reasoning (CBR) is an artificial intelligence approach. It uses previous cases to solve new, unseen and different problems. It has been widely used for enhancing decision support systems in a wide range of domains, among them in medicine and particularly in disease diagnosis [1], [2]. The CBR cycle may be described by four steps: retrieve the similar cases, reuse the information of these cases, revise the proposed solution and retain the new case. A CBR system can guarantee that it retrieves the k cases that are maximally similar to a new problem by computing the similarity of the target problem to every case in memory. The CBR methodology has also the advantage of being an easy approach to use better than knowledge model, since it avoids

the difficulties of modeling the experts' knowledge. So, CBR is also considered as a methodology and widely used in modeling decision support in different domains. However, this approach has shown some limits in relation to different methods used in its reasoning steps.

Nowadays, a major trend seems to be the extending of CBR to new reasoning tasks. This trend opened new perspectives particularly by exploring information retrieval techniques and computing and reasoning methods [2].

Section 2 is devoted to the main features of CBR and clustering. The section 3 is devoted to explain our contribution. In section 4 we give a survey of most important related work showing particularly the integration of CBR with other techniques that have contributed in decision support. We continue by presenting our approach in section 5. In section 6, we will present the application of our approach in a medical domain. In section 7, we present the experimentation and evaluate the results and finally in section 8, we give the conclusion which summarizes the paper and point out possible trends.

2. Literature Review

2.1. CBR Methodology

It uses the principle of reusing past situations to solve a current case. It is conventionally based on four tasks: retrieve, reuse, revise and retain as shown in Figure 1, according to Aamodt and Plaza [3]:

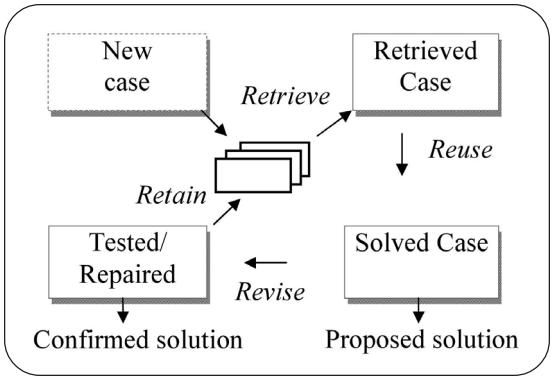

Figure 1. The CBR cycle, adapted from Aamodt and Plaza

While comparing the different tasks of CBR, the retrieval task appears as the crucial operation of the whole cycle and presents some shortcomings [1]. Several works were conducted to tackle these shortcomings and others by using suitable solutions with the aim to impact positively the CBR process [4], [5], [6].

2.2. Clustering for Retrieving Information to Support CBR

The clustering is the process of organizing objects into groups. It is a widely known data mining technique that separates objects into meaningful groups (clusters) that share common characteristics, often according to some defined distance measure. It constitutes a decision support in various sectors [7]. Typical operations of a clustering process are shown in Figure 2:

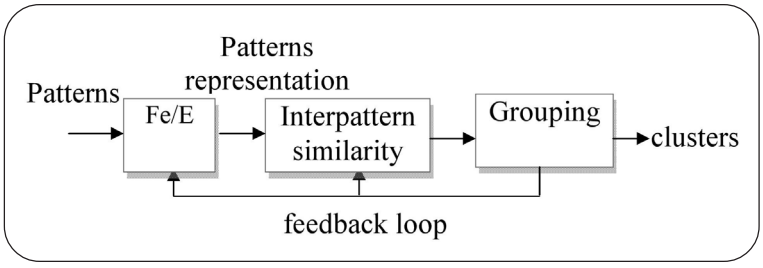

Figure 2. The standard clustering process (Fe/E: Feature Selection/Extraction)

The most popular clustering algorithm in the literature is k-means. It proceeds in four steps [7]:

1. Choose randomly k cases to form k clusters.

2. (re) Assign each case O to cluster C_i with the centre M_i such as $dist\ (O, M_i)$ is minimal.

3. Recalculate M_i for each cluster (barycenter).

4. Go to step 2, if an assignment is just made.

In the literature, several studies aimed at combining clustering with CBR have been conducted to influence the results of the retrieval [8], [9], [10], [11].

3. Contribution

In the present work, we try to experiment a new approach by using collaboration between CBR and clustering to propose an available strategy at retrieval task which permits choosing the best solution from a set of similar cases founded through a clustering of a case base.

Our aim is to focus on a "search area" (by Clustering), instead of having of bulk data which will make the adaptation phase (of CBR) very complicated and arduous. Thus, our contributions are:

• Firstly, remedy the drawbacks related to a huge set of similar cases retrieved by the classical CBR's retrieval, and consequently several solutions.

• Secondly, lighten the rest of the CBR process (search for the final solution) particularly the adaptation which is very complicated and arduous in presence of several elected cases.

• Thirdly, propose a decision support model based on a hybrid reasoning (CBR and Clustering).

4. Related Work

CBR is well suited for integration with other reasoning methodologies and has been largely deployed in multi-modal reasoning systems. The interest in multi-modal approaches involving CBR dated back to more than many years ago, and is recently increasing [5]. In particular, CBR has proved to be well suited for integrations regarding to a vast amount of works that have been carried out. In this section, we will present some of these works according to the most productive search axes.

4.1. Integrations with Rule-Based Reasoning (RBR)

As it is well known, RBR consists of using rules through a backward or forward reasoning, exploiting the data to make a decision. RBR was the first approach to be successfully integrated with CBR. As an efficient tool, RBR was well integrated with CBR and many solutions were proposed. Verma et al. [12] proposed a hybrid solution by using a framework model based on data mining (rules) and CBR. It consists of a knowledge base, CBR and a data mining subsystem, to propose recommended actions to users by a knowledge driven model. This combination aims to increase the ability to solve problems and improve suggestion accuracy. Cabrera and Edye [13] used an integration of rules and CBR to diagnose clinical case of acute bacterial meningitis. They proposed a framework which is initially applied to the pre-diagnosis stage with a basic diagnostic rules and if the pre-diagnosis stage is successful, then there is a solution to the problem which is presented to the user, offering the possibility of repairing the new case and if the case is not obvious or simple, the pre-diagnosis is not applicable and the system proceeds the case using the CBR method. Saraiva et al. [14] applied RBR to improve the CBR's retrieval process. They used symptoms, signs and personal information from patients as inputs to a model, and applied RBR to define the case's attribute weights, used in the global similarity function, and let CBR converge to the best solution. The model's output presents the probability of the patient having a type of cancer.

4.2. Integrations with Model-Based Reasoning (MBR)

MBR is an approach in which general knowledge is represented by formalizing the mathematical or physical relationships present in a problem domain. It was successfully combined to CBR in many domains. The integration CBR-MBR generally makes the adaptation process easy, improve performance and efficiency.

CASEY [2] was the first CBR-MBR system. It uses a patient case-base and a physiological model of the human heart to diagnose heart failures. It was interfaced with a previously existing MBR heart failure program and used a physiological model of the heart to match new cases to old ones and derive new diagnoses from old ones. When CASEY could not find a close enough match in its case-base, it invoked the original MBR system to solve the problem. PROTOS [2], another early CBR-MBR system, used a multi-relational model of the knowledge to diagnose auditory diseases and improve case retrieval. CARMA [2] is also a CBR-MBR decision support system used for providing advises to patients care ranchers to assist in the management of grasshopper infestations. It combines numeric models developed by entomologists with historic cases of infestations, while CBR is used to select a similar case, and the model assists in adapting that case's prediction.

4.3. Integrations with Data Mining (DM)

Data mining techniques were also used in different manners to facilitate the case-based reasoning [15], [16].

Zhuang et al. [17] integrated data mining and case-based reasoning for decision support for Pathology Ordering by General Practitioners. They used comprehensive data collected by professional pathology companies to extract knowledge from these data through data mining and give it for solving new cases of pathology test ordering problem. This system facilitated more informed evidential decision making by physicians in the area of pathology ordering. Schmidt et al. [18] suggested clustering cases into prototypes and remove redundant ones to avoid an infinite growth of a case-base, the retrieval searches only among these prototypes. This solution can simplify the adaptation task. Balakrishnan et al. [19] proposed a retinopathy prediction system based on association rules using Apriori algorithm, and case-based reasoning. The association rules are used to analyze patterns in the data set and to calculate retinopathy probability, whereas CBR is used to retrieve similar cases. This technique addresses the problem of a case-base maintenance by developing a new technique, the association-based case reduction technique (ACRT), to reduce the size of the case-base in order to enhance the efficiency while maintaining or even improving the accuracy of the CBR.

4.4. Integrations with Multi-Criteria Analysis (MCA)

Multi-Criteria Analysis has contributed to solve some limitations of CBR and many studies were conducted in this direction.

Armaghan and Renaud [20] used the integration CBR-MCA to study diabetes. This study deals with the "Retrieve" operation by using the multi-criteria decisions concept in the problem description to search for the solution in a case-based scenario. They propose using knowledge acquisition as a basis for seeking solutions from non-compensatory multi-criteria decision aids. Li and Sun [21] combined CBR and MCA to enhance a data mining process for improving detection of disease. Malekpoor et al. [22] proposed a TOPSIS-CBR approach. Initially, CBR is used to retrieve relevant cases from the database. Thereafter, inferred cases are evaluated using TOPSIS (Technique for Order Preference by Similarity to Ideal Solution: a multi-criteria decision-making technique) to prescribe an optimal dose plan. Erjaee et al. [23] proposed a specific method based on multi-criteria to propose a decision for an efficient treatment for Helicobacter pylori infection among children. Bouhana et al. [24] integrated CBR and AHP method for itinerary search.

5. The Proposed Approach

Our approach illustrated by Figure 3, is based on reducing the search space and instead of searching on the whole case base by a massive retrieval of similar cases that is the classic recipe of CBR reasoning. Indeed it does not effectively serve to collect all the neighbor cases, but it should rather focus on a small area of cases which are really closest. There are two main tasks:

1. Reduce the search space area to contain only potential cases.
2. Retrieve the best solution from this reduced area.

Based on this objective, the task (1) is performed by a clustering operation which makes 2 groups: "Candidate Cases" CC and "Not Candidate Cases" (NCC).

5.1. The Proposed Decision Support Model

We propose a medical decision support model, defined as a complete action plan which includes a set of tasks in order to ensure the main operations which help in making appropriate decision. The main operations of the decision support model are illustrated in Figure 3, which shows schematically the process from the information acquisition about the subject of decision till the final

solution of the new situation proposed to resolution:

1. Gather information and definition of a new problem.
2. Identify and select relevant descriptors to involve in clustering.
3. Initiate clustering.
4. Evaluate clusters.
5. Accept or review the proposed solution.
6. Initiate CBR.
7. Evaluate final solution deduced.
8. Memorize the current problem with its solution or drop it.

5.1.1. Clustering

Thus, instead of a massive retrieval of cases that is the classic recipe of reasoning, there is a focusing on a particular search space. Based on this principle, firstly, we compute a reduced search space (similar cases) by using specific signs and secondly, we retrieve the most interesting solution for the current case being processed. The reduction strategy impacts enormously the CBR's retrieval step. It can clearly make retrieval computationally better and hopefully more meaningful, since only cases taken under similar circumstances (specific signs) are very closest cases. So, we use a grouping procedure called "*Clustering for Retrieval*", summarized as follows:

1. Insert the current case in the case base.
2. Initialize *k* to 2 (*).
3. Cluster the case-base into k sets: "candidate cases" (CC) that share the same specific signs (**) and "not candidate cases" (NCC) which is not fairly similar.

(*) **Remark:** So, the decision maker initializes *k* to 2, as to have only two clusters. He can also repeat the test with respect to *k* > 2 to reduce more and more the cases search space and the current case will be also automatically in the cluster "Candidate Cases" (CC).

(**) **Remark:** Close in terms of values for the same specific signs.

5.1.2. CBR

It consists in finding the *n* closest cases to the proposed new problem by using a similarity measure. We used the *K*-nn method for the simplicity of its implementation. The process will begin by extracting the preliminary relevant solutions that have been considered for the *n* similar cases. These preliminary solutions are considered to determine the relevant solution (Best_Solution). Thus, CBR performs its five steps below to refine the final solution:

1. Initialize *k* to 1 (***);
2. Retrieve cases which are similar to the problem description.
3. Reuse a solution suggested by the retrieved cases.
4. Revise or adapt the solution to better fit the new case.
5. Retain the new solution once it has been validated.

(***) Remark: So, as to have only 1 closest case to the current problem. If the decision maker wants to have more closest cases, he can increase *k* with values > 1.

6. The Medical Application

The medical situation was defined previously [25], and placed usually the physician in front of a situation and will have to

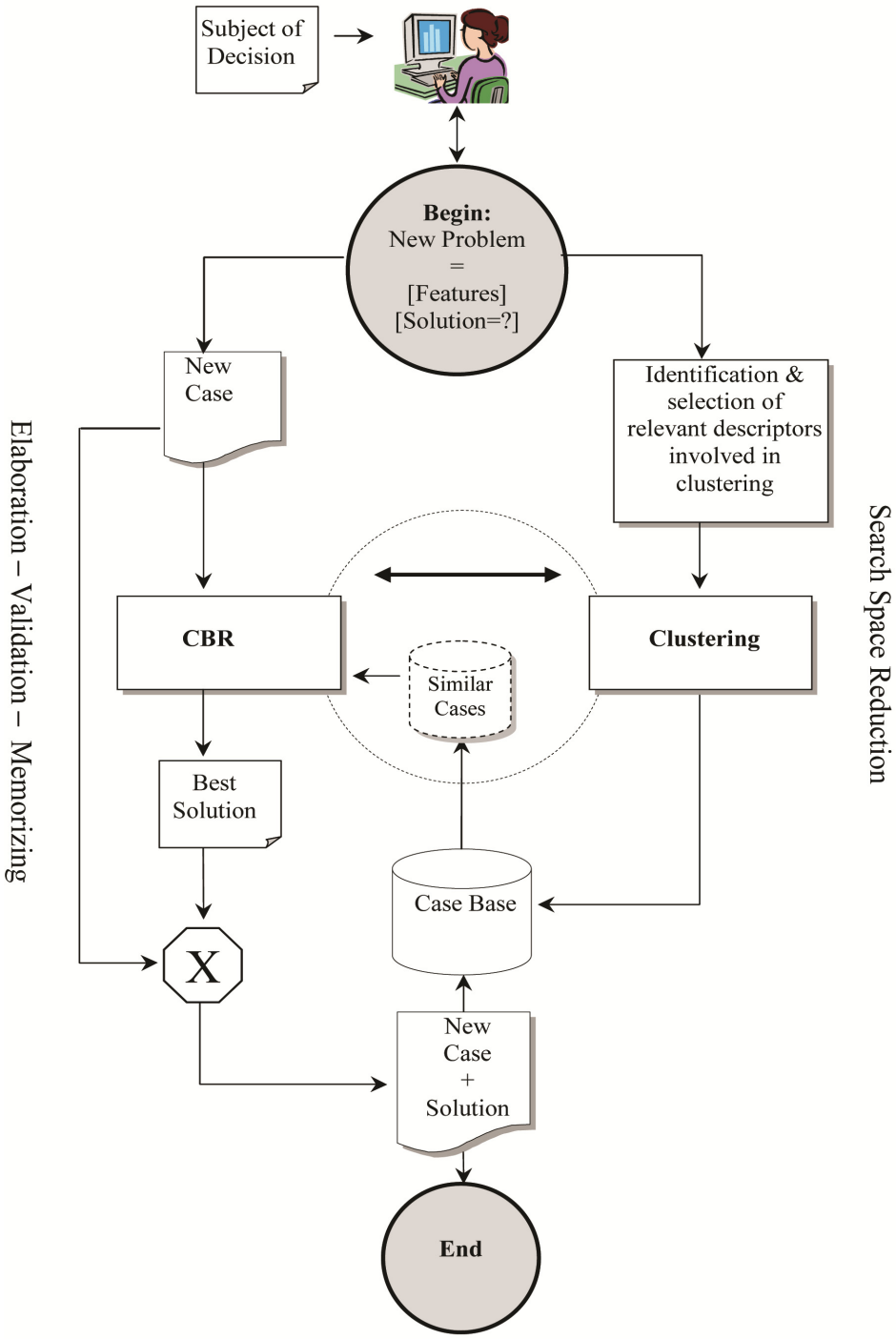

Figure 3. General overview of the Decision Support Model

explore the possible diagnosis to prescribe the best therapy. This situation is characterized by a problem definition through a set of symptoms, a set of specific signs and an exhaustive survey of possible diagnoses. The specific signs can indicate the disease's circumstances and helps to point a desired therapy, which will be more or less compliant, e.g., an elderly patient may be less compliant with a salt diet. So, the physician defines his medical situation with a set of symptoms and circumstances described by specific signs. So, the medical situation becomes a current medical case composed by *v* specific signs, *u* symptoms and a considered diagnosis [25] as in Figure 4:

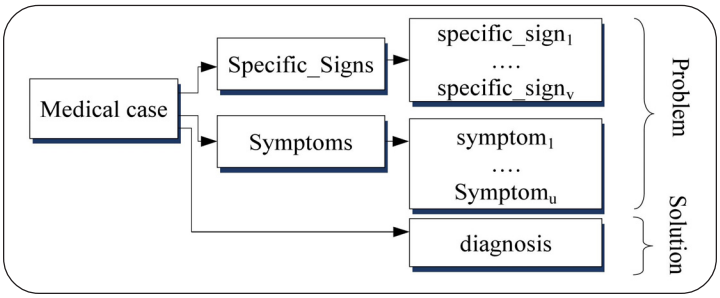

Figure 4. The medical case

The medical decision support model proposed stems from the medical situation described below (Figure 5). It is derived from the Decision Support Model in Figure 3, and with a general context stated as follows: problem definition more or less complete, an exhaustive survey of possible specific signs and symptoms.

So, to search for a solution (diagnosis), the physician must follow the steps below:

1. Gather information about the medical situation described by specific signs and symptoms.

2. Search (*) for a similar case and propose its diagnosis.

3. Evaluate the proposed diagnosis.

4. Accept the diagnosis or review the medical situation.

(*) Particularity of Clustering when managing medical situations.

Selecting relevant cases by a simple verification on attributes proves to be weakly argumentative because medical situations are not simply evaluated and stated by the simple equality operator. Many situations which are not equal but relatively similar can be dropped. Thus, specific signs appear as the most important elements in the adopted reasoning, as they will guide the clustering. So, they are the determinant elements in each medical situation before considering the other attributes (i.e., symptoms).

6.1. Clustering

This operation is necessarily preceded by a preprocessing which consists in verifying and preparing data or other specific treatment. The physician has also to check the specific signs that he wants to involve in clustering, i.e., that he considers significant with regard to the current situation. Otherwise, all the specific signs will be considered automatically. Then, an insert of the current case in the case-base is done, to be considered in the batch of cases. These basic checks completed, the grouping can start and will be handled by the following pseudo-algorithm.

Pseudo-algorithm. Clustering-For-Retrieval

```
Input: New_Case(v Specific_Sign, u Symptom, O), Case_Base
Output: Best_Cluster
Begin
Initialize k = 2 for k-means
Insert_momentarily (New_Case, Case_Base)
Features_Selection(v Specific_Sign, s_Specific_Sign)
k-means_Clustering (k, Case_Base, s_Specific_Sign, CC, NCC)
Accept_or_Refuse(CC)
If Accept CC Then go to 10
      Else return to step 6 or go to 11
   End if
Best_Cluster:=CC
End
```

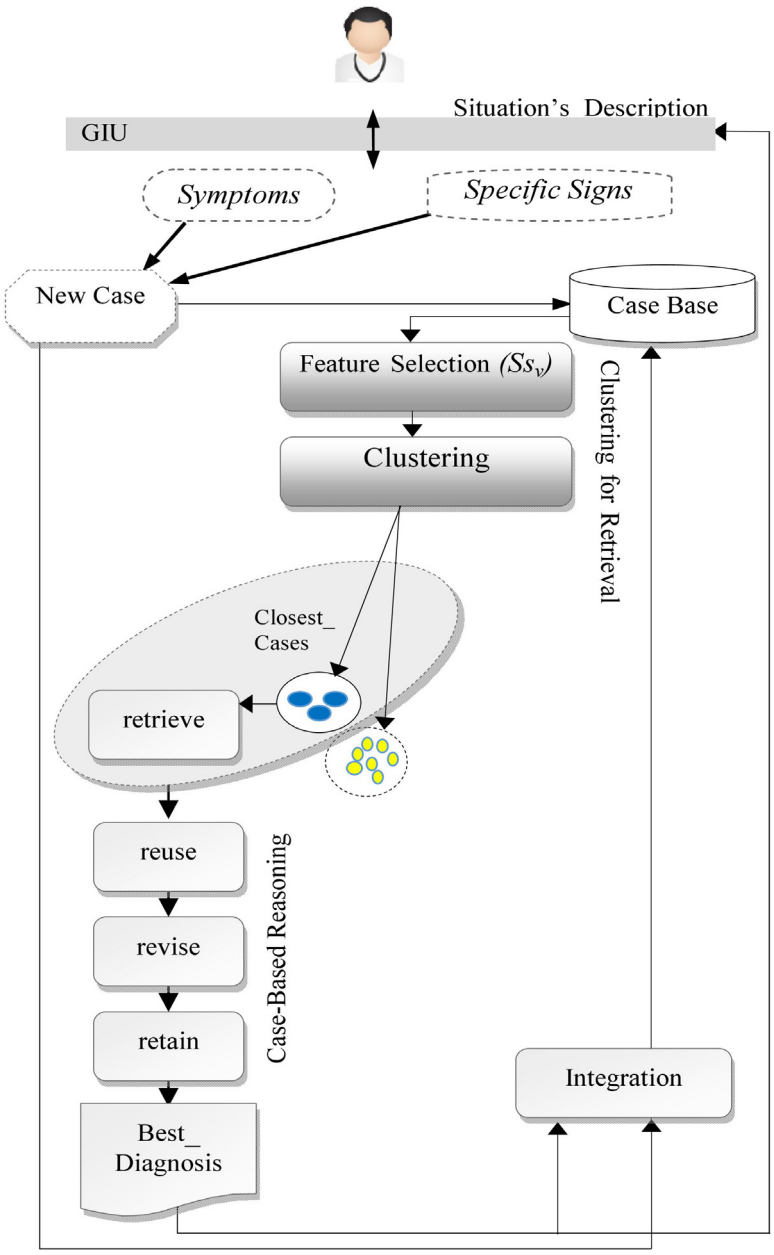

Figure 5. The multi-modal reasoning for managing a medical decision support

6.2. CBR

After receiving the best cluster, CBR initiates its first step by retrieving the most nearest neighbor cases using the *k-nn* method. It consists in finding the *n* nearest cases of the proposed case by using a similarity measure. The process will select nearest cases from the "Candidate Cases" (CC) issued from the clustering and will extract the preliminary diagnosis *(Best_Diagnosis)* from the *n* similar cases considered by the candidate cluster. The entire process will be handled by the following CBR pseudo-algorithm.

Retrieve

This operation is carried out by the *k-nn* algorithm using the Euclidean distance for quantitative attributes, to compute the similarity. It is well-suited for the data types of our case-base and does not require a long data preprocessing. Thus, by using this algorithm, the similarity measure considers only the symptoms that have been checked, through the survey carried out by

the physician. He considers them as most influential or important or enough relevant for his case's definition. For our study, with $k = 1$, we will have a 1-nearest neighbor procedure to select the best similar case and so the *"Best_Diagnosis"*.

Reuse
An evaluation of similarity between the current and the selected cases is done to determine if an adaptation is necessary or the retrieved solution can be reused directly.

Revise
It begins by adapting the retrieved solution to become a solution to the current medical problem then followed by a physician's validation to finish this step.

Retain
The physician can retain the new case with its solution and a new experienced case is stored.

Pseudo-algorithm. *CBR*

```
1: Input: Best_Cluster, New_Case(v Specific_Signs, u Symptoms, O)
2: Output: New_Case(v Specific_Signs, u Symptoms, Best_Diagnosis)
3: Begin
4: Initialize k=1 for k-nn
5: Check relevant symptoms
6: If Best_Cluster=∅ Then Diagnosis:= physician_proposed_diagnosis()
      Else
            Retrieve (Best_Cluster, K-nn, Nearest_Case)
            Diagnosis:= Nearest_Case(diagnosis)
      End If
7: Reuse(Diagnosis)
8: Revise(Diagnosis)
9: Accept_or_Refuse_Diagnosis(Diagnosis)
10: If Accept Diagnosis Then
            Best_Diagnosis:=Diagnosis
            go to 11
      Else
            Return to 5 or 8 or go to 12
      End If
11: New_Case(v Specific_Signs, u Symptoms, Best_Diagnosis)
12: End
```

7. Implementation and Experimentation

Experiments are conducted on a framework implemented in JAVA with a collaboration procedure between WEKA [26] and JColibri [27] frameworks. The framework is essentially based on the model described by the overview presented in Figure 6, where the clustering is firstly done under WEKA and secondly CBR is initiated under JColibri.

7.1. The Considered Medical Domain

"Backbone, or spine, is made up of 26 bone discs called vertebrae. The vertebrae protect your spinal cord and allow you to stand

and bend. A number of problems can change the structure of the spine or damage the vertebrae and surrounding tissue. They include:

• Infections.

• Injuries.

• Tumors.

• Conditions, such as ankylosing spondylitis and scoliosis.

• Bone changes that come with age, such as spinal stenosis and herniated disks.

Spinal diseases often cause pain when bone changes put pressure on the spinal cord or nerves. They can also limit movement. Treatments differ by disease, but sometimes they include back braces and surgery.

Diagnosis and Tests [28].

• Computed tomography (CT) – Spine.

• Discography.

• Electromyography (EMG).

• Magnetic resonance imaging (MRI) - Spine".

To describe any anomaly of the spine and determine disease, the physician uses six biomechanical features orthopedic patients: pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius grade of spondylolisthesis, with information deduced from radiographs or other X-rays image. The Figure 6 gives an illustration of these features.

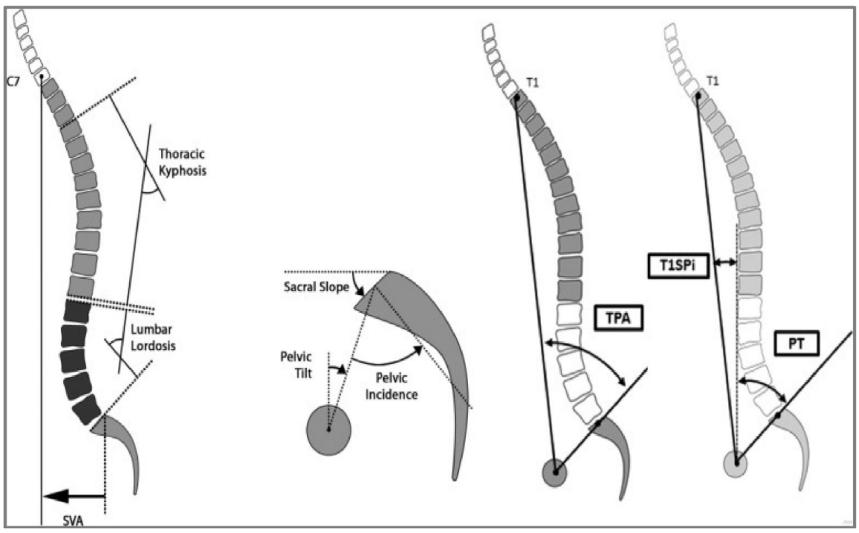

Figure 6. Illustration of the measurement of spinopelvic parameters. *SVA* = sagittal vertical axis, *TPA* = *T*1 pelvic angle, *T*1*SPi* = *T*1 spinopelvic incidence, and *PT* = pelvic tilt [29]

Considering these diseases of the spine as a field of application, we applied our proposed approach to the medical datasets "Vertebral Column Data Set of orthopaedic patients" [30].

7.2. Data Description

The proposed approach has been applied to the presumptive diagnosis of diseases of the orthopedic patients dataset which contains values for six biomechanical features used to classify orthopedic patients into 3 classes: normal, disk hernia or spondilolysthesis. The dataset was downloaded from the presumptive diagnosis of diseases of the orthopedic patient database as the raw data, from UCI Machine Learning Repository [30].

63.0278175, 22.55258597, 39.60911701, 40.47523153, 98.67291675, -0.254399986, Hernia
……
75.64973136,19.33979889,64.14868477,56.30993248,95.9036288, 69.55130292, Spondylolisthesis
……

Figure 7. Overview of data set sample of presumptive diagnosis of diseases of orthopaedic patients

7.3. Experimentation Setup

For the purposes of our experimentation, we have used all presumptive diagnosis of diseases of orthopedic patients' dataset as existing in the source and summarized as follows: 60 patients with "Disk Hernia", 150 patients with "Spondylolisthesis" and 100 patients "Normal". Afterwards, we created a dataset structure (Table 1) and transformed the presumptive diagnosis of diseases of orthopedic patients' dataset into a case-base named Ω_N (Table 2). It contains *n* cases labeled, ω_1, ω_2, …., ω_n, where each patient *i* is transformed into a case ω_i and described by the set of descriptors listed in Table 1. For each case ω_i, it's associated a target attribute noted *Y*, which has values in the set {*H, S, N*} corresponding respectively to "Disk Hernia", "Spondylolisthesis" and "Normal".

Descriptor	Label
X_1	Pelvic incidence
X_2	Pelvic tilt
X_3	Lumbar lordosis angle
X_4	Sacral slope
X_5	Pelvic radius
X_6	Grade of spondylolisthesis
Y	Diagnosis

Table 1. Case-base descriptors (structure)

ω	$X_1(\omega)$	$X_2(\omega)$	$X_3(\omega)$	$X_4(\omega)$	$X_5(\omega)$	$X_6(\omega)$	$Y(\omega)$
ω_1	63.03	22.55	39.61	40.48	98.67	-0.25	H
…							
ω_i…	89.68	32.7	83.13	56.98	129.96	92.03	S
ω_{310}							

Table 2. Overview of dataset sample of Ω_N

7.3.1. Splitting and Growing $\Omega_{HT}, \Omega_{ST}, \Omega_N$

The data are preprocessed in order to constitute the case-base Ω. It contains 60 patients with diagnosis "Disk Hernia" labeled Ω_H, 150 patients with diagnosis "Spondylolisthesis" labeled Ω_S and 100 patients with the diagnosis "Normal" labeled Ω_N. In a second step, we create Ω_L (learning case-base) and the testing case-bases Ω_{HT} and Ω_{ST}, by a random subsampling operation as follows: picking randomly 1/4. of cases from the concerned casebase, i.e., Ω_H or Ω_S and complete the rest randomly by the same

amount of cases from Ω_N in order to constitute a mixed testing case-base (sick and healthy persons):

Ω_L = 80% of Ω = 248 patients

Ω_{HT}= 1/4 of Ω_H + same amount of cases from Ω_N = 15 "*H*" + 15 "*N*" = 30 patients

Ω_{ST}= 1/4 of Ω_S + same amount of cases from Ω_N = 37 "*S*" + 37 "*N*" = 74 patients

We will have the following sizes of the different case-bases as in Table 3:

Case-Base Ω	Ω_H	Ω_S	Ω_N	Ω_L(80%)	Testing (20%)	
					Ω_{HT}	Ω_{ST}
310	60	150	100	248	30	74

Table 3. Partial case-bases

7.4. Experimentation

The experimentation is done in each case-base separately. Each case from testing base (Ω_{HT} or Ω_{ST}) is presented to the system to obtain the corresponding diagnosis from the learning case-base Ω_L. This diagnosis is compared to the one provided by the case being experienced to determine whether the system hit or failed. Thus, we use the following heuristic:

$$\forall \omega_i \in (\Omega_{HT} \text{ or } \Omega_{ST}) \text{ and } \forall \omega j \in \Omega_L \begin{cases} \text{if } Y(X(\omega i) = Y(X(\omega i)) & \text{Then "Positive"} \\ \text{Else "Negative"} \end{cases} \quad (1)$$

The heuristic (1) permits the verification and indicates if the case being processed is correctly or incorrectly identified (diagnosis) by the system. The errors are used to evaluate the overall framework performance.

7.5. Evaluation and Discussion

7.5.1. Models Evaluation

To evaluate a classifier performance, many methods are used as holdout, random subsampling, cross-validation and bootstrap [31]. However, performance measures can be used to analyze predictive models. They are based on four values of the confusion matrix as summarized in Figure 8: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Furthermore, on another verification level, the performance evaluation of the whole testing is estimated by the error rate, and it will be estimated trough common error measures. So, we can use sensitivity, specificity, accuracy and accuracy as they are defined below:

• **Sensitivity**. The percentage of patients with the disease who are correctly identified as sick.

• **Specificity**. The percentage of patients who are correctly identified as healthy.

• **Positive Predictive Value** (PPV) (aka Precision). The percentage of patients tested positive and are Truly Positive.

• **Negative Predictive Value** (NPV). The percentage of patients tested Negative and are Truly Negative.

• **Accuracy**. The percentage of patients that are correctly tested (Truly Positive or Truly Negative). It's defined by: (TP+ TN) / (TP + FP + TN + FN).

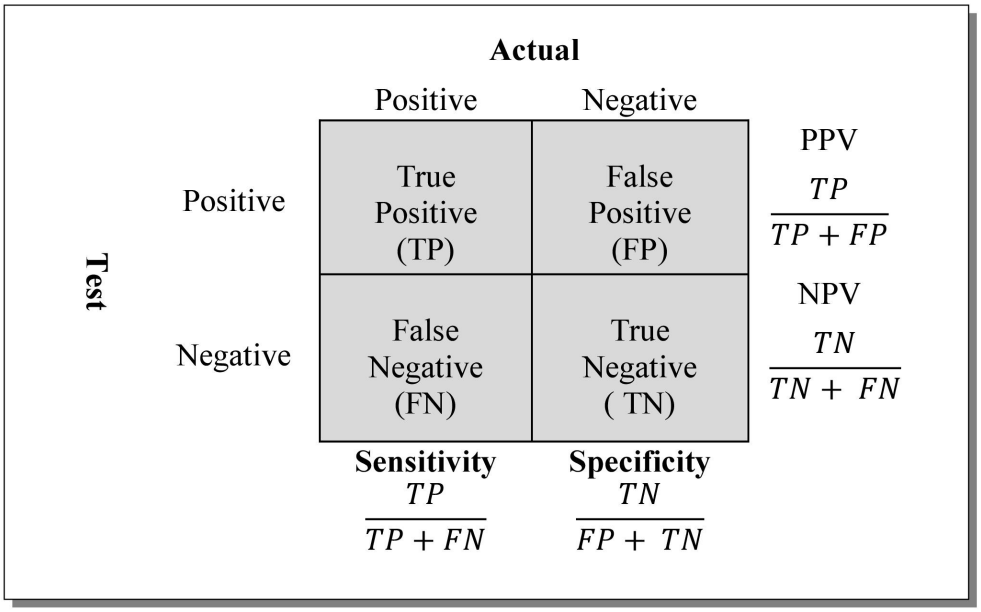

Figure 8. The normal confusion matrix 2 x 2 dimensions

7.5.2. Evaluation

We use the sensitivity measure, because it gives an idea about the recognition percentage of patients who are really identified as sick, and specificity gives an idea about the recognition percentage of patients who are really identified as healthy. We use also, PPV and NPV because they give an idea about the recognition percentage of truly positive values or truly negative values. Thereafter, at the end of all comparisons and with the formulas presented in Figure 8, we calculate the rate for sensitivity, specificity, PPV, and NPV, according to confusion matrix. The same experimental protocol is conducted respectively with testing bases Ω_{HT} and Ω_{ST}. All these measures help to evaluate the diagnostic efficiency proposed by our approach and assess its reliability. The results (in percentage) are summarized as follows:

Ω_{HT}	H+	H-	Sensitivity	Specificity	PPV	NPV	Accuracy
T+	11	2	73.33	86.66	84.61	76.47	80
T-	4	13					

Table 4. Results for diagnosis "Disk Hernia"

Ω_{ST}	H+	H-	Sensitivity	Specificity	PPV	NPV	Accuracy
T+	28	6	75.67	83.78	82.35	77.50	79.72
T-	9	31					

Table 5. Results for diagnosis "Spondylolisthesis"

T+: Tested "Positive" T-: Tested "Negative"

H+: With "Disk Hernia" H-: Without "Disk Hernia"

S+: With "Spondylolisthesis" S-: Without "Spondylolisthesis"

As sensitivity indicates how accurately our system identifies cases that have a disease or among cases with disease how often is the test right. We notice in Table 4 and Table 5 that the sensitivity is relatively high (>73 %) which indicates that our approach tend to recognize correctly a case which has "Disk Hernia" or "Spondylolisthesis".

As well as for specificity which indicates how accurately our approach identifies cases that do not have a disease or among people who are healthy how often is the test right. In Table 4 and Table 5, we note a specificity greater than 83 %, which indicates that our approach tends to recognize well cases without diseases "Disk Hernia" or "Spondylolisthesis".

On another side, the PPV indicates our system's ability to detect the presence of disease is very high (>82 %), and also for the NPV which indicates the ability to detect the absence of disease.

According to results in Table 4 and Table 5, we note that sensitivity rates, specificity, PPV and NPV are relatively high, which indicates that our system provides results (tests) close to the reality as declared in testing case-base ©T, particularly for True Positive and Negative values which indicates that our approach tends to recognize and make a good matching of diagnosis either for "Disk Hernia" or "Spondylolisthesis".

This is consolidated by accuracy which gives the percentage of cases that are correctly classified (Positive or Negative), with a value which is approximately 80 % in both cases. It indicates a relatively good matching for the diagnosis of "Disk Hernia" or "Spondylolisthesis". These observations denote that our approach shows a level of performance close to a physician's experienced diagnosis.

8. Conclusion and Future Trends

This paper aims at presenting a comprehensive view in the development and deployment of various integrations of CBR with other methodologies. Hybrid models perform both the computation and reasoning processes notably for diagnosis in medical area. CBR's integrations with other reasoning approaches continue to expand, providing both practical benefit and insight into multi-modal reasoning processes. Hybridization is fast becoming the standard, rather than the exception for CBR systems, due to user expectations as well as to complementary technical advantage of RBR, CBR and MBR.

In our study, we addressed the theoretical basis of an approach that tends to solve a problem of reasoning in the CBR methodology. First, we want to test the clustering as a principle in our approach and tested it in a medical context, with each type of diagnosis taken individually.

Later, and as a first step, we will use a single testing case-base and a single learning case-base, each one containing all types of diagnoses and will guide our experimenting with different values of k to refine and determine the most optimal value to adopt for *kmeans* algorithm. In a secondary step, we intend to evolve our approach in another orientation by using various data types and testing different similarity measures. Thus, experiment will help to refine an appropriate similarity measure, to use in situations described by a variety of attributes.

References

[1] Begum, S., Ahmed, M., Funk, P., Xiong, N., Folke, M. (2011). Case-Based Reasoning Systems in the Health Sciences: A Survey of Recent Trends and Developments. *IEEE Transactions on systems, man, and cybernetics* part c: applications and reviews, 41 (4) 421-434.

[2] Marling, C., Sqalli, M., Rissland, E., Munoz-Avila, H., Aha, D. (2002). Case-Based Reasoning Integrations. *AI Magazine*, 23 (1) 69.

[3] Aamodt, A., Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI communications, 7 (1) 39-59.

[4] Liu, C. H., Chen, L. S., Hsu, C. C. (2008). An association-based case reduction technique for case-based reasoning. *Information Sciences*, 178 (17) 3347- 3355.

[5] Bichindaritz, I., Marling, C. (2010). Case-based reasoning in the health sciences: Foundations and research directions. *In: Computational Intelligence in Healthcare* 4. Springer Berlin Heidelberg, 127-157.

[6] Marling, C., Rissland, E., Aamodt, A. (2005). Integrations with case-based reasoning. *The Knowledge Engineering Review*, 20 (3) 241-245.

[7] Kesavaraj, G., Sukumaran, S. (2013). A study on classification techniques in data mining. *In*: Computing, Communications and Networking Technologies, Fourth International Conference on. IEEE, p. 1-7.

[8] Neshat, M., Sargolzaei, M., Toosi, Nadjaran., A., Masoumi, A. (2012). Hepatitis disease diagnosis using hybrid case based reasoning and particle swarm optimization. *ISRN Artificial Intelligence*.

[9] Yang, Q., Wu, J. (2000). Keep it simple: A case-base maintenance policy based on clustering and information theory. *In*: Conference of the Canadian Society for Computational Studies of Intelligence. Springer, Berlin, Heidelberg, 102-114.

[10] Vong, C. M., Wong, P. K., Fai, Weng. Ip (2010). Case-based classification system with clustering for automotive engine spark ignition diagnosis. *In Computer and Information Science* (*ICIS*), 2010 IEEE/ACIS 9th International Conference on. IEEE, 17-22.

[11] Uddin, Mobyen., A., Banaee, H., Loutfi, A. (2013). Health monitoring for elderly: An application using case-based reasoning and cluster analysis. *ISRN Artificial Intelligence*.

[12] Verma, L., Srinivasan, S., Sapra, V. (2014). Integration of rule-based and case-based reasoning system to support decision making issues and Challenges. *In Intelligent Computing Technics* (*ICICT*), International Conference on. IEEE, p. 106-108.

[13] Cabrera, M. M., Edye, E. O. Integration of rule-based expert systems and case-based reasoning in an acute bacterial meningitis clinical decision support system, arXiv preprint arXiv:1003.1493[Online]. Available: https://arxiv.org/ftp/arxiv/papers/1003/1003.1493.pdf

[14] Saraiva, R., Perkusich, M., Silva, L., Siebra, C., Perkusich, A. (2016). Early diagnosis of gastrointestinal cancer by using case-based and rule-based reasoning. *Expert Systems with Applications*, 61, 192-202.

[15] Bichindaritz, I. (2015). Data Mining Methods for Case-Based Reasoning in Health Sciences. *In*: ICCBR (Workshops), 184-198.

[16] Yuan, G., Hu, J., Yinghong, P. (2011). Research on CBR system based on data mining. *Applied Soft Computing*, 11 (8) 5006–5014.

[17] Zhuang, Z. Y., Churilov, L., Burstein, F., Sikaris, K. (2009). Combining Data Mining and Case-based Reasoning for Intelligent Decision Support for Pathology Ordering by General Practitioners. *European Journal of Operational Research*, 195(3) 662- 675.

[18] Schmidt, R., Montani, S., Bellazzi, R. (2001). Cased-based reasoning for medical knowledge-based systems. *International Journal of Medical Informatics*, 64 (2) 355-367.

[19] Balakrishnan, V., Shakouri, M. R., Hoodeh, H. (2012). Integrating association rules and case-based reasoning to predict retinopathy. *Maejo International Journal of Science and Technology*, 6 (3) 334-343.

[20] Armaghan, N., Renaud, J. (2012). An application of multi-criteria decision aids models for Case-Based Reasoning. *Information Sciences*, 210, p. 55-66.

[21] Li, H., Sun, J. (2009). Hybridizing principles of the Electre method with case-based reasoning for data mining: Electre-CBR-I and Electre-CBR-II. *European Journal of Operational Research*, 197 (1) 214-224.

[22] Malekpoor, H., Mishra, N., Sumalya, S., Kumari, S (2016). An efficient approach to radiotherapy dose planning problem: a TOPSIS case-based reasoning approach. *International Journal of Systems Science: Operations & Logistics*, p. 1-9.

[23] Erjaee, A., Bagherpour, M., Razeghi, S., Dehghani, S. M., Imanieh, M. H., Haghighat, M. (2012). A multi-criteria decision making model for treatment of Helicobacter pylori infection in children. *Hong Kong Journal of Paediatrics*, 17 (4) 237-42.

[24] Bouhana, A., Abed, M., Chabchoub, H. (2011). An integrated Case-Based Reasoning and AHP method for personalized itinerary search. Logistics, 4th International Conference on. IEEE, p. 460-467.

[25] Author, coauthor. (2016). Clustering to Enhance Case-Based Reasoning, *In*: Modelling and Implementation of Complex Systems. *Springer International Publishing*, 2016, p. 137-151.

[26] Markov, Z., Russell. I. (2006). An introduction to the WEKA data mining system. *ACM SIGCSE Bulletin,* 38 (3) 367-368.

[27] Recio-Garcia, J.A., Diaz-Agudo, B., Gonzalez-Calero, P (2008). JColibri2 Tutorial. Universidad Complutense de Madrid: Madrid, Spain.

[28] MedlinePlus. (2016). MedlinePlus [Online]. Available: https://medlineplus.gov/spineinjuriesanddisorders.html#cat78

[29] JB & JS. (2016). JB & JS [Online]. Available: http://journals.lww.com/jbjsjournal/pages/articleviewer.aspx?year=2015&issue=12020&article=00002&type=Fulltext

[30] UCI Machine learning (2016). UCI Machine learning [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Vertebral+Column#

[31] Beleites, C., Salzer, R., Sergo, V. (2013). Validation of soft classification models using partial class memberships: An extended concept of sensitivity & co. applied to grading of astrocytoma tissues. *Chemometrics and Intelligent Laboratory Systems*, 122, p. 12-22.